CLASS-DEPENDENT TWO-DIMENSIONAL LINEAR DISCRIMINANT ANALYSIS USING TWO-PASS RECOGNITION STRATEGY

Peter Viszlay, Martin Lojka, Jozef Juhár

Department of Electronics and Multimedia Communications, Technical University of Košice
Park Komenského 13, 041 20 Košice, Slovakia
Email: peter.viszlay@tuke.sk, martin.lojka@tuke.sk, jozef.juhar@tuke.sk

ABSTRACT

In this paper, we introduce a novel class-dependent extension of two-dimensional linear discriminant analysis (2DLDA) named CD-2DLDA, applied in automatic speech recognition using two-pass recognition strategy. In the first pass, the class labels of test sample are obtained using baseline recognition. The labels are then used in CD transformation of test features. In the second pass, recognition of previously transformed test samples is performed using CD-2DLDA acoustic model. The novelty of the paper lies in improvement of the present 2DLDA algorithm by its modification to more precise, class-dependent estimations repeated separately for each class. The proposed approach is evaluated in several scenarios using the TIMIT corpus in phoneme-based continuous speech recognition task. CD-2DLDA features are compared to state-of-the-art MFCCs, conventional LDA and 2DLDA features. The experimental results show that our method performs better than MFCCs and LDA. Furthermore, the results confirm that CD-2DLDA markedly outperforms the 2DLDA method.

Index Terms— class-dependent transformation, discriminant analysis, scatter matrix, time alignment

1. MOTIVATION AND BACKGROUND

Two-dimensional linear discriminant analysis (2DLDA) [1] is a popular extension of classical linear discriminant analysis (LDA). It was mainly proposed to overcome the singularity problem in LDA implicitly, by redefining the data representation model. Recently, 2DLDA has been applied in several application fields such as face recognition [1–3] and speech recognition [4, 5].

Generally, discriminant analysis methods try to find transformations that minimize the within-class scatter and maximize the between-class scatter of the training data. The transformation can be computed in class-independent (CI) or class-dependent (CD) manner. In CI approach, one global transformation matrix is determined, which is used to transform the data. This is the most used approach in linear transformations such as LDA or PCA (Principal Component Analysis). In CD approach, contrary to the class-common methods, one transformation matrix is determined for each class separately. Different transformation spaces are constructed for different classes [6]. CD transformations usually provide more precise projections of the samples because the transformation matrices are estimated with strong respect to the discrimination information contained in the corresponding class. In this way, the samples are projected to several spaces, instead of one global identical space constructed for all classes, which satisfies only for the statistical majority of samples.

Several class-dependent extensions of linear transformations were proposed in the past such as class-dependent PCA to robust feature extraction [7]. Authors in [8] designed class-dependent LDA for robust speech recognition or feature extraction in face recognition [6]. However, 2DLDA has not been extended to class-dependent approach yet. This fact motivated us to explore this approach, especially in speech recognition. We were also partially motivated by our previous work [9], where we investigated the performance of 2DLDA. In this paper, we modify classical 2DLDA proposed by authors in [1] and introduce its new class-dependent extension. As in other class-dependent methods, the fundamental issue is the classification phase, where class labels have to be assigned to test samples. In order to do this, we use two-pass recognition strategy. In the first pass, time aligned phoneme sequences are converted into labels that are used as input to CD-2DLDA transformation. In the second pass, the final recognition is performed on the transformed test sample with appropriate CD-2DLDA based acoustic model. Our results confirm that CD-2DLDA achieves higher recognition accuracies compared to 2DLDA.

The rest of this paper is organised as follows. In Section 2, classical LDA and 2DLDA methods are reviewed. Section 3 gives the detailed description of proposed CD-2DLDA method. The experimental setup is given in Section 4. The experimental evaluation and discussion is presented in Section 5. Finally, the results are concluded in Section 6.

The research presented in this paper was supported by the Research and Development Operational Program funded by the ERDF under the projects ITMS-26220220155 (50%) and ITMS-26220220182 (50%).

2. LDA AND 2DLDA ESTIMATION

2.1. Classical LDA

LDA is a well-known dimensionality reduction and transformation method used in automatic speech recognition. It maps the N-dimensional input data to p-dimensional subspace ($p < N$) while retaining maximum discrimination information. The aim of LDA is to find a transformation matrix $W \in \mathbf{R}^{N \times p}$ that projects each vector $\boldsymbol{x}_i$ to vector $\boldsymbol{y}_i$ as $\boldsymbol{y}_i = W^T \boldsymbol{x}_i$. In case of class-independent LDA, the within-class scatter matrix S_W and the between-class scatter matrix S_B are defined as:

$$S_W = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in \Pi_i} (\boldsymbol{x} - \mu_i)(\boldsymbol{x} - \mu_i)^T, \tag{1}$$

$$S_B = \sum_{i=1}^{k} (\mu_i - \mu)(\mu_i - \mu)^T, \tag{2}$$

where $\mu_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \Pi_i} \boldsymbol{x}$ are the class mean vectors and $\mu = \frac{1}{n} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in \Pi_i} \boldsymbol{x}$ is the global mean vector. In speech recognition, $\boldsymbol{x}$ represents a supervector created by concatenating of C basic feature vectors computed on successive speech frames. The scatter matrices are estimated from the training data partitioned into k classes Π_i, where class Π_i contains n_i elements. Notice that $n = \sum_{i=1}^{k} n_i$ is the total number of elements [4]. The transformation matrix W can be obtained by solving the generalized eigenvalue problem $S_B \boldsymbol{v} = \lambda S_W \boldsymbol{v}$, for $\lambda \neq 0$, where $\boldsymbol{v}$ and λ represent the eigenvectors and eigenvalues, respectively. W is finally obtained by eigendecomposition of the matrix $S_{WB} = S_W^{-1} S_B$.

2.2. Two-dimensional LDA

2DLDA was primarily designed to overcome the singularity or undersampled problem in classical LDA implicitly [1]. The key difference between LDA and 2DLDA is in the data representation model. Different from LDA, in which the data is represented in vector space, matrix representation is adopted by 2DLDA [4]. 2DLDA alleviates the difficult computation of the eigendecomposition of scatter matrices related to LDA. Since it works with matrices instead of high-dimensional supervectors, the eigendecomposition in 2DLDA is computed on scatter matrices with much smaller sizes than in LDA. This reduces the processing time and memory costs of 2DLDA compared to LDA [1].

2DLDA aims at finding two transformation matrices $L \in \mathbf{R}^{r \times l_1}$ and $R \in \mathbf{R}^{c \times l_2}$ to project X_j to $Y_j \in \mathbf{R}^{l_1 \times l_2}$ as $Y_j = L^T X_j R$, $j \in \langle 1; n \rangle$, where $X_j \in \mathbf{R}^{r \times c}$ are matrix represented training speech samples belonging to k classes Π_i. X_j represents a matrix composed from C concatenated acoustic vectors computed on successive speech frames [4], as in LDA.

Transformation matrices L and R can be obtained by maximizing the Fisher ratio of between-class and within-class scatter matrices after projection. The within-class and between-class scatter matrix coupled with R are defined as:

$$S_w^R = \sum_{i=1}^{k} \sum_{X \in \Pi_i} (X - M_i) R R^T (X - M_i)^T, \tag{3}$$

$$S_b^R = \sum_{i=1}^{k} n_i (M_i - M) R R^T (M_i - M)^T, \tag{4}$$

and the within-class and between-class scatter matrix coupled with L are defined as:

$$S_w^L = \sum_{i=1}^{k} \sum_{X \in \Pi_i} (X - M_i)^T L L^T (X - M_i), \tag{5}$$

$$S_b^L = \sum_{i=1}^{k} n_i (M_i - M)^T L L^T (M_i - M), \tag{6}$$

where $M_i = \frac{1}{n_i} \sum_{X \in \Pi_i} X$ is the i-th class mean matrix and $M = \frac{1}{n} \sum_{i=1}^{k} \sum_{X \in \Pi_i} X$ is the global mean matrix. Due to difficult computing of optimal L and R simultaneously, authors in [1] derived an iterative algorithm to find L and R by iteratively fixing another one. The algorithm firstly computes optimal L for fixed R using (3) and (4) by eigendecomposition of $\left(S_w^R\right)^{-1} S_b^R$. In the next step, with fixed L it computes optimal R using (5) and (6) by eigendecomposition of $\left(S_w^L\right)^{-1} S_b^L$. The iterative procedure is several times repeated. It should be noted that the sizes of scatter matrices in 2DLDA are much smaller that those in LDA. Specifically, the size of S_w^R and S_b^R is $r \times r$ and the size of S_w^L and S_b^L is $c \times c$. More detailed description can be found in [1].

3. CLASS-DEPENDENT TWO-DIMENSIONAL LDA

3.1. Description of CD-2DLDA

In this work, we were motivated by the model of class-dependent LDA, in which the within-class scatter matrix is separately computed and stored for each training class, while the between-class scatter matrix is computed in the same way as in class-independent LDA. Transformation matrix is then separately determined by eigendecomposition of scatter matrices for each class.

The key idea behind CD-2DLDA is to apply original 2DLDA algorithm separately to each class Π_i to obtain a couple of transformation matrices L_i and R_i, while the between-class matrices stay always the same. L_i and R_i are then used to transform the feature vectors with class label i. Using these assumptions, we define the class-dependent within-class scatter matrix $S_{w_i}^R$ of class i coupled with R as:

$$S_{w_i}^R = \sum_{X \in \Pi_i} (X - M_i) R R^T (X - M_i)^T \tag{7}$$

and the class-dependent within-class scatter matrix $S_{w_i}^L$ of class i coupled with L as:

$$S_{w_i}^L = \sum_{X \in \Pi_i} (X - M_i)^T L L^T (X - M_i). \quad (8)$$

The class mean matrices M_i, the global mean matrix M and between-class scatter matrices are computed in the same way as in class-independent 2DLDA (see Section 2.2). Therefore, instead of having one 2D transform (as in class independent 2DLDA), we have multiple transforms, one for each class.

3.2. Constraints and solutions

In case of CD-2DLDA (compared to CD-LDA) several problems arise. The first one is that transformation matrices L_i and R_i are computed by iterative estimation resulting from 2DLDA (see Section 2.2), while the transformation matrices in CI-LDA or CD-LDA can be computed directly, without using an iterative optimization. In order to extend 2DLDA to CD-2DLDA, similar iterative algorithm is applied to each training class Π_i to obtain $S_{w_i}^L$, $S_{w_i}^R$, L_i and R_i.

The second one is that the quality of scatter matrices $S_{w_i}^L$, $S_{w_i}^R$, S_b^L, S_b^R and transformation matrices L_i and R_i markedly depends on the number of iterations. We found out that successful convergence of CD-2DLDA can be achieved when a reasonable number of iterations is reached. We investigated the quality of CD scatter matrices and we found that few iterations are enough to convergence (the scatter and transformation matrices did not update during the next iteration).

Furthermore, the complexity of CD-2DLDA is more difficult because the algorithm considers k, $(l_1 \times l_2)$-dimensional spaces $\mathcal{L}_i \otimes \mathcal{R}_i$, which are tensor products of two spaces [1]. In other words, there are two statistical estimators coupled with L_i and another two estimators coupled with R_i, instead of two statistical estimators coupled with L or R in 2DLDA or one statistical estimator coupled with W_i in CD-LDA. In our case, the i-th class is represented by $S_{w_i}^L$, $S_{w_i}^R$, L_i and R_i, while S_b^L and S_b^R are the same for all classes.

Another limitation in the core algorithm of 2DLDA is that L directly depends on estimation of S_w^R, which is directly estimated using initial R. In addition, R then directly depends on estimation of S_w^L, which is directly estimated using L, computed previously. Due to these dependencies, matrices S_b^L and S_b^R for CD-2DLDA can not be computed directly from the data (as in CD-LDA). In other words, to estimate S_b^L and S_b^R properly, 2DLDA has to be performed before the main CD-2DLDA estimations. Note that only the training phase of 2DLDA without transformation of speech data is required.

3.3. Two-pass recognition based on CD-2DLDA

CD-2DLDA consists of two separate two-dimensional supervised transformations. The first one is the transformation of training vectors, whose class labels are known from the embedded training or forced alignment. Therefore, CD-2DLDA transform can be simply applied with corresponding L_i and R_i. After 2D transform, we used the transformed training set for HMM-based acoustic model training (see Section 4).

However, we had to modify the recognition step considerably, because the class labels of test samples are not known. In order to transform test samples in class-dependent manner, their labels are needed to be known before the transformation phase. We used two-pass recognition strategy to meet this condition. The first pass is represented by classical recognition step, in which the baseline acoustic model (see Section 4) is used to determine the most likely hypotheses for unknown speech recording in form of time aligned segments on phoneme level. The resulting phone-based time alignment is directly used to determine class labels of the current recording. These labels are then provided to supervised CD-2DLDA transformation of the recording. Subsequently, in the second recognition pass CD-2DLDA-based acoustic model is used to recognize the transformed recording finally. Note that the language resources are not changed during recognition phases.

The CD-2DLDA algorithm with two-pass recognition can be summarized in following steps (see Figure 1):

1. Compute M_i, M, S_b^L and S_b^R according to CI-2DLDA.
2. Compute CD-2DLDA parameters - for i from 1 to k, compute $S_{w_i}^R$, L_i, $S_{w_i}^L$ and R_i iteratively with known M_i, M, S_b^L and S_b^R.
3. Transform the training set with L_i and R_i according to CD-2DLDA concept.
4. Train CD-2DLDA based acoustic model.
5. Perform first recognition pass with baseline acoustic model and determine class labels of the test samples.
6. Transform the test samples using the labels with corresponding L_i and R_i according to CD-2DLDA concept.
7. Perform second recognition pass with CD-2DLDA acoustic model and recognize the transformed test samples.

4. EXPERIMENTAL SETUP AND CONDITIONS

In this work, we used the TIMIT acoustic phonetic speech corpus to train and test our ASR system. For training and testing, we used complete sets. Acoustic models were trained using the provided phone segmentation. 61 original phones were mapped to final inventory of 41 symbols.

The speech signal was preemphasized and windowed every 10ms using Hamming window of length 25ms. Fast Fourier transform and Mel filter-bank analysis with 20 channels were applied to the windowed segments. After applying discrete cosine transform (DCT), 12 Mel-frequency cepstral coefficients (MFCC) and the 0-th coefficient were retained.

In LDA and 2DLDA processing, 13-dimensional MFCC vectors were used as input features. We used supervectors

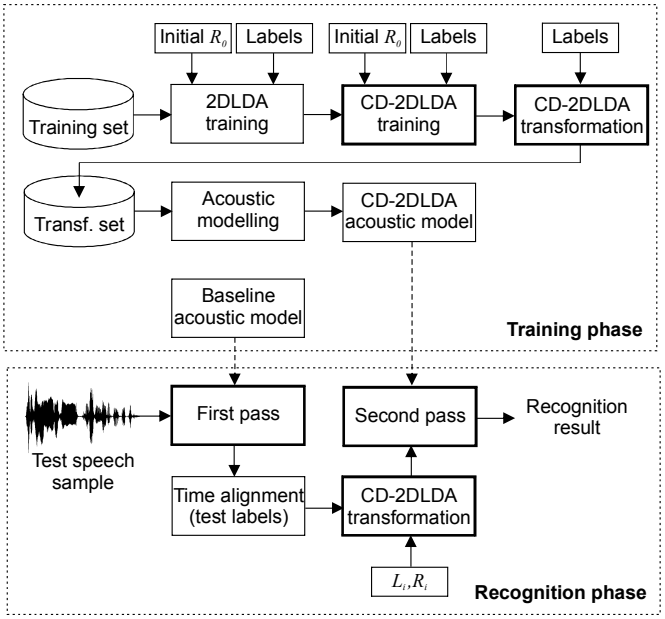

Fig. 1. Block diagram of the training and testing (recognition) phase based on CD-2DLDA features.

in LDA with $C = 3$. In 2DLDA and CD-2DLDA we used 2D tokens composed from three basic vectors ($C = 3$), i.e. $r \times c = 13 \times 3$. We performed dimension reduction. In LDA, $N = 39$ LDA coeffs. were reduced to $p = 13$. In 2DLDA and CD-2DLDA, 2D tokens were reduced to vectors with $l_1 = 13$ and $l_2 = 1$. In order to compare the baseline and LDA, 2DLDA and CD-2DLDA ASR system regularly, 13 LDA (2DLDA, CD-2DLDA) features were retained after transformation and expanded with Δ and $\Delta\Delta$ coefficients. The acoustic models of baseline, LDA, 2DLDA and CD-2DLDA system had the same dimension. The training class labels were obtained from the phonetic alignment. The number of classes k corresponded to 41 English phones contained in the corpus for all discriminant analyses.

The ASR system used context-independent monophones modelled using three-state left-to-right Hidden Markov Models (HMMs) on phone level. The number of Gaussian mixtures per state was a power of 2, starting from 1 to 32. The number of monophone models corresponded to the number of phonemes and LDA, 2DLDA and CD-2DLDA classes. The recognition network was configured as a word network not as a phone network. Therefore, the vocabulary size was approx. 6000 words. For testing, a word lattice was created from a bigram language model [10]. HMM training, testing and evaluation by HTK Toolkit [11] were performed.

In order to evaluate the experiments we chose the word-level recognition accuracy computed as $Acc. = \frac{H-I}{N} \times 100\%$, where H is the number correctly recognized words, I is the number of insertions and N is the total number of labels [11].

Mixtures	MFCC	LDA	2DLDA	CD-2DLDA	FA
1	54.08	56.60	56.65	**57.30**	56.51
2	55.89	56.67	56.79	**57.88**	57.40
4	58.37	58.74	58.85	**60.41**	59.69
8	59.60	60.56	60.62	**61.75**	60.53
16	61.00	61.75	61.77	**62.90**	61.55
32	62.61	62.65	62.89	**63.84**	62.45

Table 1. Comparison of different speech recognition systems.

5. EXPERIMENTAL EVALUATION AND RESULTS

In this section, the proposed class-dependent approach to 2DLDA is experimentally evaluated and the results are being presented. Table 1 gives the main comparison of recognition accuracy of five ASR systems with different types of features:

1. *Baseline ASR system*: conventional 39-dim. MFCCs;
2. *Conventional LDA*: 39-dim. supervector composed from three successive frames is reduced to 13-dimensional LDA features and expanded with Δ and $\Delta\Delta$ coeffs.;
3. *2DLDA*: similarly as in LDA, the feature matrix of dimension 13×3 is reduced to 2DLDA vector of dimension 13×1 and expanded with Δ and $\Delta\Delta$ coeffs.;
4. *CD-2DLDA*: equally as in 2DLDA;
5. *FA*: equally as in 2DLDA; specific system used as alignment reference. It was used to evaluate CD-2DLDA on test class labels resulting from the forced alignment (FA) carried out using ortographic transcriptions and baseline acoustic model. One of the goals in the experimental work was to investigate the performance of CD-2DLDA based on test labels obtained from the forced alignment.

From the Table 1 it can be seen that compared with the baseline, LDA clearly improved the recognition performance. Further, 2DLDA slightly improved the LDA performance for all mixtures with the same model dimension, as was expected. The most important results are listed in the column marked as "CD-2DLDA". Note that the listed results are the maximum values obtained from CD-2DLDA transformations using three different time alignments (from the first recognition pass). It is clear that CD-2DLDA consistently performed better than 2DLDA for all Gaussian mixtures. The maximum absolute improvement of 2DLDA achieved by CD-2DLDA is $+1.56\%$ at 4 mixtures. The presented results are also graphically compared in Figure 2.

In our experiments, we tested CD-2DLDA on several time alignments resulting from the first recognition pass. During comprehensive testing we found that the best results are produced by CD-2DLDA using alignments resulting from recognition based on acoustic models with 32, 64 and 128 mixtures. Note that we have to differentiate the number of mixtures used in the main evaluation ($1 - 32$) and the number of mixtures used in the first pass to generate the time alignment

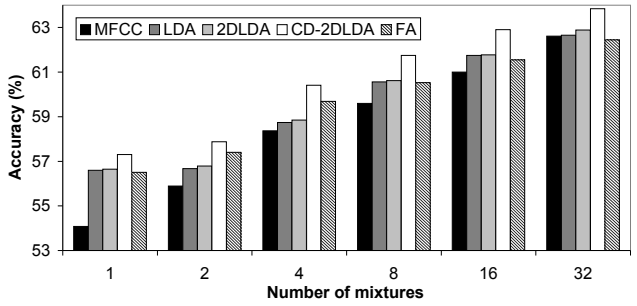

Fig. 2. Maximum performances of different ASR systems.

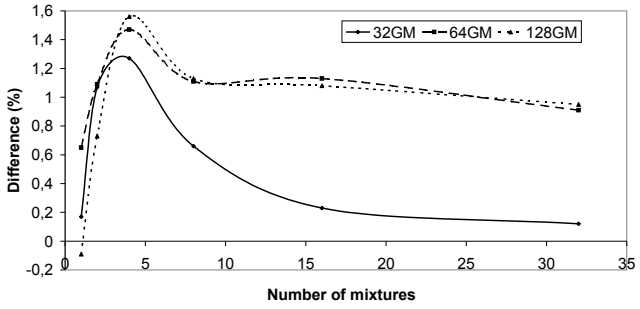

Fig. 3. Absolute improvements (%) of 2DLDA depending on the number of mixtures used in time alignment in first pass.

$(32-128)$. The comparison absolute improvement of 2DLDA by CD-2DLDA depending on the number of mixtures used to generate the time alignment is given in Figure 3. It is clear that the most effective time alignments are the ones resulting from the first pass with 64 and 128 mixtures. On average, the absolute improvements are alternating around $+1\%$.

Finally, from the table it can be seen that CD-2DLDA based on FA did not perform very well. We expected that FA will perform best because the time alignment was generated using the reference transcriptions. We suppose that the alignment based on the correct transcription may not always be automatically the best way to obtain the test labels. In real applications, the FA system does not have any meaning because the correct reference alignment is not available. Therefore, the result of this test does not have a big significance.

The number of iterations in 2DLDA and CD-2DLDA estimation was equal to $I = 10$. All recognition phases were performed with equal word insertion log probability ($p = -8.0$), except the time alignment phases ($p = -30.0$).

6. CONCLUSION

In this paper we introduced a novel class-dependent approach to 2DLDA using two-pass recognition concept. We have proven that the proposed method performs better than 2DLDA. Another upcoming direction in research related to CD-2DLDA is to investigate its performance at higher number of mixtures. We want to apply it on more real-life corpora, such as COSINE, etc. We would like to apply it also in Slovak large vocabulary continuous speech recognition system.

REFERENCES

[1] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1569–1576, 2005.

[2] D. Luo, Ch. Ding, and H. Huang, "Symmetric two dimensional linear discriminant analysis (2DLDA)," in *Proc. of the Intl. Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 2820–2827.

[3] J-Y. Gan, S-B. He, and B. Luo, "Face recognition based on two-dimensional heteroscedastic discriminant analysis," in *Proc. of the Intl. Conf. on Information Science and Engineering*, 2009, pp. 852–856.

[4] X. B. Li and D. O'Shaughnessy, "Clustering-based Two-Dimensional Linear Discriminant Analysis for Speech Recognition," in *Proc. of the Annual Conf. of the International Speech Communication Association*, 2007, pp. 1126–1129.

[5] S.-B. Chen, Y. Hu, B. Luo, and R-H. Wang, "Heteroscedastic discriminant analysis with two-dimensional constraints," in *Proc. of ICASSP'92*, 2008, pp. 4701–4704.

[6] J. Liang, "Class-dependent LDA for feature extraction and recognition," in *Proc. of the Intl. Conf. on Computer Science and Information Technology*, 2010, pp. 614–618.

[7] H. Abbasian, B. A. Nasersharif, and A. Akbari, "Genetic programming based optimization of class-dependent PCA for extracting robust MFCC," in *Proc. of INTERSPEECH'08*, 2008, pp. 1541–1544.

[8] H. Abbasian, B. A. Nasersharif, A. Akbari, and M. Rahmani, "Optimized linear discriminant analysis for extracting robust speech features," in *Proc. of the Intl. Symposium on Communications, Control and Signal Processing*, 2008, pp. 819–824.

[9] J. Juhár and P. Viszlay, "Linear feature transformations in Slovak phoneme-based continuous speech recognition," in *Modern Speech Recognition Approaches with Case Studies*. 2012, pp. 131–154, InTech Open Access.

[10] J. Juhár, J. Staš, and D. Hládek, "Recent progress in development of language model for Slovak large vocabulary continuous speech recognition," in *New Technologies - Trends, Innovations and Research*. 2012, pp. 261–276, InTech Open Access.

[11] S. Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University, 2006.